Standard for Sharing Clinical Study Data (v1.0)

Draft for Public Comment

Version Date: 26 February 2024

Review Notes:

- This document is a draft for public comment.
- Please make comments using the comment submission form: https://members.crdsalliance.org/document/dl/421. Comments that do not use the form will not be considered.
- Email your completed form to the Clinical Research Data Sharing Alliance (CRDSA) secondary use standards work group at su@members.crdsalliance.org
- The closing date for comments is 31 May 2024.

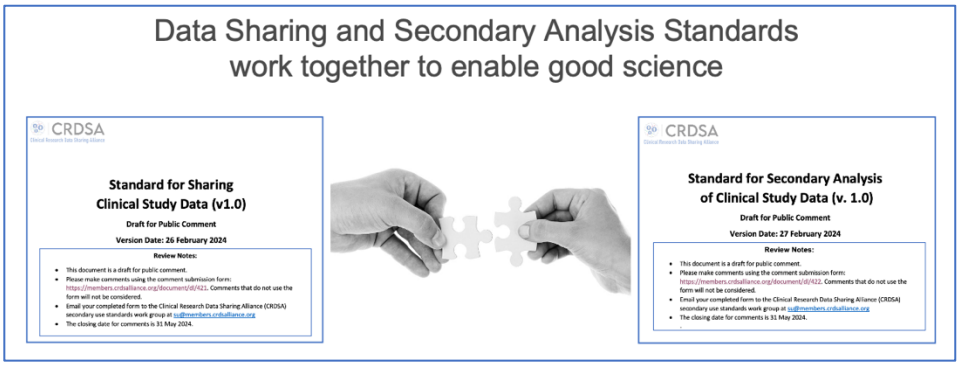
.

CRDSA is a multi-stakeholder alliance that serves the clinical data-sharing ecosystem. Our mission is to accelerate the discovery and delivery of life-saving and life-changing therapies to patients by expanding the research value of secondary use data. Broad access to these data has the power to transform the research process, improve trial design and delivery, and benefit the patients who donate their time and their data as part of the clinical development process.

PREFACE

The Clinical Research Data Sharing Alliance has created two draft documents outlining standards for both the sharing and secondary analysis use of clinical study data. Both standards aim to facilitate the responsible sharing and use of anonymized individual patient data (IPD) from clinical studies to enable further research and scientific understanding while protecting patient privacy and innovation.

Each document applies to a different audience - broadly, contributors of clinical trial data and researchers using that data. However, it is important to recognize that the standards are complementary and intended to work together to facilitate good science. For example, the standard for secondary analysis is predicated on adherence to the data sharing standard, as the former relies on the proper sharing of data, metadata, and documents outlined in the latter.



Each standard provides principles, supporting criteria, and best practices for clinical study data sharing policies and procedures. CRDSA considers the principles and supporting criteria to be mandatory. However, it is recognized that adherence is not possible or applicable in some cases. Each document provides a checklist so that implementing organizations can allow for deviations.

The standards can be adopted by data sharing platforms, funders, research institutions, and scientific journals. Implementation may vary depending on the organization and its use case(s) and include adopting the standards as written or modifying them to suit organizational needs (provided alterations are clearly outlined).

When public comments are incorporated and final standards are published, the two standards establish consistent guidelines for responsibly sharing clinical study data and conducting robust secondary analyses of that data to advance scientific knowledge while safeguarding key considerations.

CONTENTS

1. INTRODUCTION

Why do we need a standard for sharing clinical study data?

Sharing anonymized individual patient data (IPD) from clinical studies provides opportunities to conduct secondary research to verify the original research findings, test new hypotheses, and further scientific understanding while respecting the expectations of research participants who donate their data for scientific use [1].

Achieving widespread research use of clinical study IPD requires data to be shared outside the original researchers' institution or research collaboration — that is, external data sharing — while retaining as much research utility as possible. In addition, for this data to be used effectively, before the data is accessed, there needs to be information available (e.g., metadata) so researchers can assess whether data from the study is likely to be relevant for their research question. The data needs to be shared in usable formats and provided with study documents and information so researchers can understand and navigate the data [2]. These considerations for sharing data to maximize utility should be balanced with the need to:

1. Protect the privacy of patients and those involved in the research
2. Protect innovation and intellectual property
3. Ensure efficient use of resources

The need to maximize data utility while protecting innovation and privacy in cost-efficient ways has led to variability in data sharing policies [3] that determine which studies are shared, what data and documents are shared, and when and how they are shared. For example, different types of datasets and documents may be shared, and they may be shared at different times in the process. There may also be differences in the anonymization approaches to protect privacy and intellectual property, which in turn may be dependent on the method of access used (such

as whether the data is shared openly or shared under highly secure controlled access conditions) [4].

These differences can negatively affect the value of this data for secondary research, particularly where the secondary research seeks to use data from multiple studies that are shared in different ways.

This standard is informed by a survey of clinical trial data users, which provided insight into what data and documents provide value for researchers and what study metadata and documents should be shared prior to data requests and data access [2].

This standard can enable clinical study data to be shared in more consistent ways that maximize utility while protecting innovation and privacy. This standard can also create process efficiency and information transparency that will benefit the research community and, equally important, benefit data contributors by ensuring that their investment in data preparation time and resources will maximize research outcomes.

This standard can be adopted and required by research funders and sponsors and used to develop their clinical study data sharing policies and procedures. The standard may also be used by others to assess whether clinical study data is being shared responsibly and in ways that benefit science.

Scope of this standard

This initial version of the standard addresses sharing IPD from interventional clinical studies conducted in patients and non-interventional clinical studies using patient data.

Sharing data from secondary analyses (e.g., analysis-ready datasets for meta-analyses) is out of the current scope.

Organization of this standard

This standard provides principles, supporting criteria, and best practices for clinical study data sharing policies and procedures. CRDSA considers the principles to be mandatory. Where needed, the principles are supplemented with criteria to be followed to meet the principle. Non-mandatory guidance is provided as best practices, which are described with “should” terminology.

These principles, criteria, and best practices are not intended to provide step-by-step instructions; rather, they are intended to be a framework for clinical study data sharing that can be adapted to different circumstances as appropriate.

The principles and criteria may not be applicable in every circumstance, and a checklist is provided where any deviations from the principles can be explained.

2. PRINCIPLES AND PRACTICES

2.1 DATASETS

PRINCIPLE: ANONYMIZED RAW DATASETS AND ANALYSIS-READY DATASETS ARE TO BE MADE AVAILABLE FOR DATA SHARING

Sharing the anonymized raw dataset (e.g., the data collected for each patient in a clinical trial) maximizes data utility because it can be used for a wide range of analyses beyond the scope of the original study. It can also be used to verify the transformations used to create the analysis-ready dataset.

Sharing the anonymized, analysis-ready dataset allows other researchers to reproduce the results and reproduce the findings of the original study. Sharing this dataset saves other researchers the time and resources required to derive the analysis endpoints, and provides insight into how the derivations were programmed, including assumptions for missing or inconsistent data points. It enables them to focus on conducting further analyses or exploring different research questions without having to create an analysis-ready dataset from the raw data.

Best Practices

- Anonymized raw datasets and analysis-ready datasets from interventional clinical trials should be shared in Study Data Tabulation Model (SDTM) [5] and Analysis Data Model (ADaM) [6] data schema, respectively, because these models provide a standardized way to organize and structure clinical trial data. Doing so helps enable consistency across different studies, making it easier to compare and combine data from various sources.
- Datasets should be shared using widely available statistical analysis software file types (e.g., datasets created using R statistical software). File types requiring software licenses should be shared through the use of open transport protocols (e.g., .xpt). The use of delimited flat files (e.g., .csv) should be avoided.
- The raw and analysis-ready datasets do not contain original radiographs, electrocardiograms, and the like; information derived from these sources may be included in clinical study datasets. If a researcher requires these materials for analysis, the original researchers should try to provide them if they are readily available and patient privacy can be protected [7]. If these materials are not available, this should be made clear when making the study available for data sharing.

2.2 SUPPORTING DOCUMENTATION AND METADATA

PRINCIPLE: SUPPORTING DOCUMENTS AND METADATA ARE TO BE SHARED SO THAT RESEARCHERS CAN UNDERSTAND AND USE THE DATASETS. THE FOLLOWING ARE TO BE INCLUDED IN THE DATA CONTRIBUTION:

2.2.1 Published information and metadata References, identification numbers, or links to primary publications and at least one study registration are to be made available. Publicly available information such as publications and registrations provide summary-level information that can help researchers understand the datasets available. For example, they provide start and end dates, study location, study design, study population, inclusion and exclusion criteria, treatments, primary and secondary outcomes, number of patients included, adverse events, summary results, and interpretations. Study registrations may also provide access to the data sharing plan, statistical analysis plan, and study protocol.

2.2.2 Latest study protocol or plan (including details of any amendments). This document describes the objectives, design, methodology, statistical considerations, and organization of a clinical study. Sharing the anonymized (or redacted) final study protocol or plan allows other researchers to understand how the original study was designed, conducted, and analyzed. This protocol/plan may be made publicly available through study registration.

2.2.3 Dataset specification. A dataset specification for a clinical study is the metadata that describes the datasets, such as variable labels, variable descriptions, code lists, and data formats. Providing the dataset specification allows other researchers to understand how the datasets are organized and managed.

2.2.4 Data dictionary. The data dictionary defines data types, formats, value definitions and variables as well as variable level transformations, data elements modified into standard or custom models, and terminologies and their meaning.

2.2.5 Annotated case report form (aCRF). This is a blank case report form with descriptions of the data collected and how they are mapped in the raw dataset. Sharing this information helps other researchers better understand the dataset and its context.

2.2.6 Statistical analysis plan (SAP). The SAP outlines the prespecified statistical methods and analyses planned for the study. Sharing the SAP enables other researchers to understand how the data was analyzed. It can also help other researchers replicate analyses in the original study, thus ensuring that they are interpreting the data correctly.

2.2.7 Clinical study report (CSR). A CSR is written for studies sponsored by the biopharmaceutical industry. This document provides a comprehensive summary of the study, including detailed data on the methods, results, and conclusions. Sharing the CSR allows other researchers to access relevant information about the study. Before this document is shared, it is anonymized or redacted to protect privacy. Other information may also be redacted to protect commercially confidential information (e.g., see Health Canada guidance [8]). The core CSR (without patient level listings) is to be shared — sharing the CSR synopsis is insufficient to meet this criterion.

2.2.8 Encoding information. Encoding is the process of converting or representing data collected in a clinical study using a specific coding system or standard terminology. For example, the Medical Dictionary for Regulatory Activities (MedDRA) [9] is a widely used coding system for standardizing the representation of adverse events in clinical trial data. Where clinical study data has been encoded and the details are not included in the minimum standard document set or in the raw (SDTM) dataset, data contributors are to provide specific encoding information or references (e.g., for adverse events, concomitant medications).

Best Practices

Further information that could be shared with the datasets includes:

- **Analysis Data Reviewer's Guide.** This is written for studies sponsored by the biopharmaceutical industry. It provides guidance and instructions to reviewers (such as regulators) who are tasked with reviewing and validating the statistical analyses conducted for a clinical trial. As with the SAP, other researchers can use this document to replicate analyses in the original study, thus ensuring that they are interpreting the data correctly.
- **The Study Reviewer's Guide.** This is written for studies sponsored by the biopharmaceutical industry. It contains detailed information on the data elements, data collection procedures, data validation rules, and data quality checks that reviewers should perform during the review process. It outlines the steps and criteria for data review and may include specific guidelines for resolving any discrepancies. It helps reviewers understand the context and background of the study, enabling them to identify potential data issues and ensure that the data is of high quality and suitable for analysis.
- **Analytic code.** This is the computer code used to carry out analyses in the original study. Sharing the code enables other researchers to replicate the findings and understand the coding methods used. Sharing analytic code also enables other researchers to build upon the code to refine methods and conduct additional analyses more efficiently.

2.3 DATA TRANSFORMATION REPORT

PRINCIPLE: DATA TRANSFORMATIONS ARE TO BE DOCUMENTED IN A STUDY-SPECIFIC TRANSFORMATION REPORT

2.3.1 Anonymization methodology: For other researchers to understand how the data has been anonymized and which information has been changed or removed to protect privacy, the anonymization and redaction methods used are to be clearly documented and made available with study datasets and documents. Information supplied is to include the following:

- Specificity on the risk assessment; application of quantitative or qualitative methodology; and the relevant factors considered in the assessment
- References to the anonymization methods used
- The applicable regulatory guidance followed

2.3.2 Dataset transformation: A dataset-level transformation report is to be provided. Examples of transformation and transparency tools and approaches can be found in Annex 1. The report is to include the following:

- **Variables:** information on any variables that have been redacted or changed
- **Adverse events:** information on any changes to adverse events, inclusive of any redactions or reclassifications (e.g., to a high-level group term or MedDRA [9] system organ class)
- **Data removal:** information on any dataset domains or data types (e.g., genetic data, exploratory biomarkers) that have been removed

2.3.3 Transformation report format: The transformation report is to contain the dataset domain, variable name, the applicable change or transformation made to the variable, and the reason for the action taken. To illustrate, a transformation report may be in the following format:

Dataset Domain	Variable Name	Change/Transformation
Domain Examples: - Demographics (DM) - Concomitant Medications (CM) - Adverse Events (AE)	Variable Examples: - AGE - SEX - RACEO - CMTRT - COTXT	Name the action taken. For example: - Values removed/suppressed/dropped - Values offset/shifted by [technique name] - Outliers grouped (example: top-to-bottom coding) - Values grouped to higher granularity or per [standard/specification reference] - Values (numerical) generalized by [specify the parameters, banding intervals, etc.]

		Explain or reference to a reasoning underpinning the decision to take such action, such as: - sensitive patient information - variable blank - grouped or banded to reduce reidentification risk (e.g., country to region)

Best Practices

- **Adverse Event (AE), Concomitant Medications (CM), and Medical History (MH) domains or equivalent datasets:** These domains contain crucial information enabling intervention safety assessments. The anonymization approach should seek to promote end-user utility by retaining as much detail as possible. In particular, adverse event records should be retained. If there is a valid reason for removal (e.g., to protect patient privacy), detailed information should be provided to disclose the level of adverse event coding removed and explain any potential impacts on secondary analysis. This approach applies to similar domains with important information, such as Concomitant Medications and Medical History.

- **Risk-based anonymization:** If compatible with regulatory guidance, a risk-based data anonymization method should be used because these approaches help to effectively balance research utility with the need to protect privacy. These methods take into account the level of privacy risk and commonly use measures of the risk (or probability) of reidentification [4] [10].

- **Other anonymization methods:** Where quantitative risk-based approaches are not used, other (e.g., rule-based) peer reviewed methods and best practices should be used to anonymize data. However, it should be recognized that the specifics of each study, such as the study disease (e.g., common or rare disease), the sensitivity of the data collected, the level of granularity or detail, and how the data is shared, may mean that the same anonymization methods may not be equally effective (e.g., a higher level of anonymization may be needed for open access vs. controlled access models).

2.4 PROVISION OF SUPPORTING DOCUMENTATION

PRINCIPLE: SUPPORTING DOCUMENTATION IS TO BE MADE AVAILABLE TO RESEARCHERS INDEPENDENT OF DATA REQUEST OR DATA ACCESS

Providing researchers with supporting documents and metadata in advance of access to or provision of individual patient data (IPD) helps researchers determine whether a study is likely to include data relevant for their research question before they request or access data. This information can be made available on request and/or it can be made publicly available.

Providing this information in advance promotes efficiency in the data sharing process, allowing researchers to make informed decisions about which studies to access based on the likely relevance to their research objectives. This can save time and resources for data contributors and researchers.

2.4.1 The following supporting documents and metadata are to be made available publicly or on request in advance of providing IPD:

- Summary protocol (e.g., study protocol registrations)
- Study results (e.g., primary publications and at least one study result registration)
- Latest study protocol or plan (including details of any amendments)
- Statistical analysis plan

2.4.2 For studies that have been processed for secondary use (e.g., where IPD has been shared), the following supporting documents are to be made available to researchers as soon as the study is processed or available, independent of data request or data access:

- Annotated case report form (aCRF)
- Core clinical study report (see 2.2)
- Data transformation report (see 2.3)
- Dataset specification
- Data dictionary
- Encoding information

3 CHECKLIST

Any deviations from the principles can be explained in the checklist below.

NOTE: For a principle to be deemed satisfied, there must be documented and verifiable support for compliance.

		Yes/No/NA	NA Explanation
DATASETS	Are the anonymized raw dataset and analysis-ready dataset available for data sharing?		
SUPPORTING DOCUMENTATION AND METADATA	Are references, identification numbers, or links to primary publications and at least one study registration made available?		
	Is the latest study protocol or plan (with details of any amendments) shared?		
	Is the dataset specification shared?		
	Is the data dictionary shared?		
	Is the annotated case report form (aCRF). shared?		
	Is the statistical analysis plan (SAP) shared?		
	Is the core clinical study report (CSR) shared?		
	Is encoding information shared?		
DATA TRANSFORMATION REPORT	Are the anonymization and redaction methods included in a study-specific data transformation report?		
	Are dataset-level transformations included in the data transformation report?		
	Does the data transformation report format include the dataset domain, variable name, the applicable change or transformation made to the variable, and the reason for the action taken?		
PROVISION OF SUPPORTING DOCUMENTATION	Are the following documents available (publicly or by request) in advance of providing IPD? - Summary protocol (e.g., through study protocol registrations) - Study results (e.g., primary publications and at least one study result registration) - Latest study protocol or plan (with details of any amendments)		

	- Statistical analysis plan		
	For studies that have been processed for secondary use, are the following supporting documents available to researchers independent of data request or data access? - Annotated case report form (aCRF) - Core clinical study report - Data transformation report - Dataset specification - Data dictionary - Encoding information		

321

322

323

4. REFERENCES

324

325

[1] S. R. Karpen, J. K. White, A. P. Mullin, I. O'Doherty, L. D. Hudson, K. Romero, S. Sivakumaran, D. Stephenson, E. C. Turner and J. Larkindale, "Effective data sharing as a conduit for advancing medical product development.," *Ther Innov Regul Sci,* vol. 55, no. 3, pp. 591-600, May 2021.

[2] E. Odame, T. Burgess, L. Arbuckle, A. Belcin, G. Jones, P. Mesenbrink, R. Walls and A. Mann, "Establishing a Basis for Secondary Use Standards for Clinical Trials," *Applied Clinical Trials,* March 9, 2023.

[3] CRDSA Data Protection Work Group, "2023 Update – A Review of Biopharma Sponsor Data Sharing Policies and Protection Methodologies," 28 September 2023. [Online]. Available: https://crdsalliance.org/resources/#dpwg.

[4] S. Bamford, S. Lyons, L. Arbuckle and P. Chetelat, "Sharing Anonymized and Functionally Effective (SAFE) Data Standard for Safely Sharing Rich Clinical Trial Data," *Applied Clinical Trials,* vol. 31, no. 7/8, April 8, 2022.

[5] CDISC, "SDTM," [Online]. Available: https://www.cdisc.org/standards/foundational/sdtm.

[6] CDISC, "ADaM," [Online]. Available: https://www.cdisc.org/standards/foundational/adam.

[7] P. Nisen and F. Rockhold, "Access to Patient-Level Data from GlaxoSmithKline Clinical Trials.," *N Engl J Med.,* vol. 369, pp. 475-478, 2013.

[8] Health Canada, "Public Release of Clinical Information: guidance document," [Online]. Available: https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance/document.html.

[9] MedDRA, [Online]. Available: www.meddra.org.

[10] "ISO/IEC 27559:2022 Information security, cybersecurity and privacy protection – Privacy enhancing data de-identification framework," ISO, [Online]. Available: https://www.iso.org/standard/71677.html.

ANNEX 1

TransCelerate Privacy Methodology

https://www.transceleratebiopharmainc.com/initiatives/privacy-methodology-for-data-sharing-2/

TransCelerate BIOPHARMA INC.
Accelerating Answers.

TRANSPARENCY CHECKLIST TEMPLATE

The Transparency checklist should be included with each study package shared, to provide information on the data protection methods applied to facilitate use of the data.

Note: if an anonymization report has already been prepared, and it covers the elements described in the below checklist, then sharing the anonymization report rather than filling in the checklist is acceptable.

TRANSPARENCY CHECKLIST		
PART 1. Privacy Approach Applied		
1a. [MANDATORY] Specify the approach applied for each type of variable.	**Select one for each type of variable:** • **TransCelerate's Recommended Approach** • **TransCelerate's Compatible Approach** • **Other Approach**	**Please elaborate on the approach applied (e.g., whether variables were removed or altered, rationale for why certain variables were removed/altered)**
Unique Identifiers *NOTE: Where identifiers have been removed, the rationale should be described*		
Dates		
Verbatim/Free Text		
Banding of Variables		
Patient Demographics (sex, race, ethnicity)		
Data With Low Frequencies		

TRANSPARENCY CHECKLIST		
Sensitive Information		
Adverse Events & Medical History *NOTE: If any MedDRA levels are removed, please describe the reasons behind the removal.*		
Concomitant Medications *NOTE: The version information provided here unless it is provided in a variable in the dataset.*		
Geographic Location		
Records of Participants Who Have Died		
1b. [MANDATORY] If appliable, please elaborate on the approach applied to the following variables.		
Information Collected Under Copyright Licenses		
Data Derived from Genomic Data		
Seasonality		
PART 2. Data Participants in Dataset		
2a. [MANDATORY] Has any individual participant's data been removed from the dataset due to anonymization		

TRANSPARENCY CHECKLIST	
requirements? Indicate Yes/No	
2b. [MANDATORY] If yes in 2a, provide the current number of participants included in the dataset.	
2c. [OPTIONAL] If yes in 2a, provide the rationale for removal.	
PART 3. Other Information	
3. [OPTIONAL] Indicate if your anonymization report has been provided in the study upload package or if one can be publicly accessed. If available, it is strongly recommended that you share your anonymization report or equivalent document. Please redact any information identifying the vendor before sharing. *If the anonymization report covers any of the other elements in this Transparency checklist, there is no need to duplicate the information.*	
4. [OPTIONAL] Please describe the data format of the provided dataset, e.g., SDTM, ADaM and/or Other. *In DataCelerate®, indicate the data format*	

TRANSPARENCY CHECKLIST	
specifications in the Transparency Checklist.	
5. [OPTIONAL] Please indicate if the study is for an indication where seasonality is an important factor, any adaptations that were required to the methodology (e.g., how variables related to regions and dates have been protected).	
6. [OPTIONAL] Please provide any other information that is considered helpful to a future research team.	